

Non-Local Priors for High-Dimensional Estimation

David Rossell¹, Donatello Telesca²

Author's Footnote

¹ University of Warwick, Department of Statistics

² UCLA, Department of Biostatistics

February 21, 2014

Abstract

Non-local priors (NLPs) possess appealing properties for high-dimensional model choice, e.g. parsimony or consistency of posterior model probabilities. Their use for estimation has not yet been studied in detail, partially due to difficulties in characterizing the posterior on the parameter space. Here we give a general representation of NLPs as mixtures of truncated distributions. This enables simple posterior sampling and flexibly defining NLPs beyond previously proposed families. We study the linear regression case in detail by giving posterior sampling algorithms and assessing finite sample behavior in engineered data sets. These studies show low serial correlation in posterior samples and notable high-dimensional estimation with finite sample sizes. Relative to benchmark and hyper-g priors, SCAD and LASSO, NLPs combine small MSE and short posterior intervals for spurious covariates with competitive MSE and frequentist coverage for non-zero coefficients, suggesting a higher estimation efficiency. Our findings also contribute to the debate of whether different priors should be used for estimation and model selection, showing that selection priors perform remarkably well for high-dimensional estimation.

Keywords: Model Selection, MCMC, Non Local Priors

1 Introduction

The class of non-local prior (NLP) distributions has appealing properties for Bayesian hypothesis testing and variable selection. Relative to local priors, NLPs discard spurious covariates at a faster rate as the sample size n grows, while preserving exponential learning rates to detect non-zero coefficients [Johnson and Rossell, 2010]. This additional parsimony enforcement has important consequences in high-dimensional settings. In Normal regression models, when the number of variables p grows at rate $O(n^\alpha)$ with $0.5 \leq \alpha < 1$, the posterior probability $P(M_t | \mathbf{y}_n)$ assigned to the data-generating model M_t converges in probability to 1 when using NLPs. In contrast, when using local priors $P(M_t | \mathbf{y}_n)$ converges to 0 [Johnson and Rossell, 2012]. Despite these attractive properties, the use of NLPs for estimation problems has not been studied in detail yet.

To fix ideas, denote the observed outcome by $\mathbf{y}_n \in \mathcal{Y}_n$, where \mathcal{Y}_n is the sample space. For ease of discussion we assume a parametric density $f(\mathbf{y}_n | \boldsymbol{\theta})$ indexed by $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ and defined with respect to an adequate σ -algebra and σ -finite dominating measure. Suppose we entertain a collection of K models M_1, \dots, M_K with corresponding parameter spaces $\Theta_1, \dots, \Theta_K \subseteq \Theta$, where $\Theta_k \cap \Theta_{k'}$ has 0 Lebesgue measure for any $k \neq k'$. We say that $\pi(\boldsymbol{\theta} | M_k)$, an absolutely continuous prior density for $\boldsymbol{\theta}$ under model M_k (and support Θ_k), is a NLP if it converges to 0 as $\boldsymbol{\theta}$ approaches any value $\boldsymbol{\theta}_0$ that would be consistent with another model $M_{k'}$.

Definition 1.1. *Let the conditions outlined above pertain and $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|$ be a suitably defined distance between $\boldsymbol{\theta}$ and an arbitrary value $\boldsymbol{\theta}_0 \in \Theta_{k'}$, $k' \neq k$. $\pi(\boldsymbol{\theta} | M_k)$ is a non-local prior under model M_k if for all $\boldsymbol{\theta}_0$, k' and $\epsilon > 0$ there exists $\eta > 0$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \eta$ implies that $\pi(\boldsymbol{\theta} | M_k) < \epsilon$.*

Intuitively, NLPs define probabilistic separation of the models under consideration, which is the basis for their improved learning rates.

Consider now an estimation problem where interest lies either on the posterior distribution conditional on a single model $\pi(\boldsymbol{\theta} | M_{k^*}, \mathbf{y}_n)$ or the model averaging posterior $\pi(\boldsymbol{\theta} | \mathbf{y}_n) = \sum_{k=1}^K \pi(\boldsymbol{\theta} | M_k, \mathbf{y}_n) P(M_k | \mathbf{y}_n)$. Let M_t be the data-generating model. Whenever $P(M_t | \mathbf{y}_n) \rightarrow 1$, the oracle property $\pi(\boldsymbol{\theta} | \mathbf{y}_n) \rightarrow \pi(\boldsymbol{\theta} | M_t, \mathbf{y}_n)$ is achieved by direct application of Slutsky's theorem. For instance, under the conditions in Johnson and Rossell [2012] for certain high-dimensional linear models only NLPs achieve $P(M_t | \mathbf{y}_n) \rightarrow 1$ and are hence guaranteed to achieve the oracle property. However, using NLPs for estimation problems presents important challenges. Strategies to compute posterior model probabilities and explore the model space are

available, but $\pi(\boldsymbol{\theta} \mid M_k, \mathbf{y}_n)$ presents multi-modalities that grow extreme in high dimensions. For instance, product MOM and iMOM priors (and corresponding posteriors) on p parameters have 2^p modes. Another difficulty lies in finding functional forms for $\pi(\boldsymbol{\theta} \mid M_k)$ that lead to simple calculations, which limits the flexibility in choosing the prior.

The main contribution of this manuscript lies in the representation of NLPs as mixtures of truncated distributions (Section 3). We prove that this is a one-to-one characterization, in the sense that any NLP can be represented as such a mixture, and any non-degenerate mixture induces a NLP. The representation provides an intuitive justification for NLPs, adds flexibility in their definition and enables efficient posterior sampling even under strong multi-modalities (Section 4). We also study finite-sample estimation performance via case studies (Section 5). Throughout we place emphasis on high-dimensional problems and devote our interest to proper priors, as their shrinkage properties outperform improper priors even in moderate dimensions. See the early arguments in Stein [1956], the review on penalized likelihood advantages over maximum likelihood estimation in Fan and Lv [2010], or related results from a Bayesian perspective in George et al. [2012].

2 Non local priors for testing and estimation

To provide intuition, we first introduce a simple example built on basic principles. In the fundamental Bayesian paradigm the prior depends only on one’s knowledge (or lack thereof), but in practice the analysis goals may affect the prior choice. For instance, different priors are often used for estimation and model selection (see Bernardo [2010] and references therein advocating the use of common priors in a decision-theoretic setup). While the practice might deviate from the fundamental paradigm, it can be argued that some preferences are hard to formalize in the utility function and may be easier to account for in the prior. For instance, in high dimensions point masses provide a convenient simplification for interpretation and computation, which may be otherwise unfeasible.

Suppose we wish to both estimate $\theta \in \mathbb{R}$ and test $H_0 : \theta = 0$ *vs.* $H_1 : \theta \neq 0$, and that the analyst is comfortable with specifying a (possibly vague) prior for the estimation problem. As an example, the gray line in Figure 1 shows a $\text{Cauchy}(0, 0.25)$ prior expressing confidence that θ is close to 0, *e.g.* $P(|\theta| > 0.25) = 0.5$ and $P(|\theta| < 3) = 0.947$. Setting a prior for the testing problem is more challenging, as the analyst wishes to assign positive prior probability to H_0 . To be consistent, she aims to preserve the

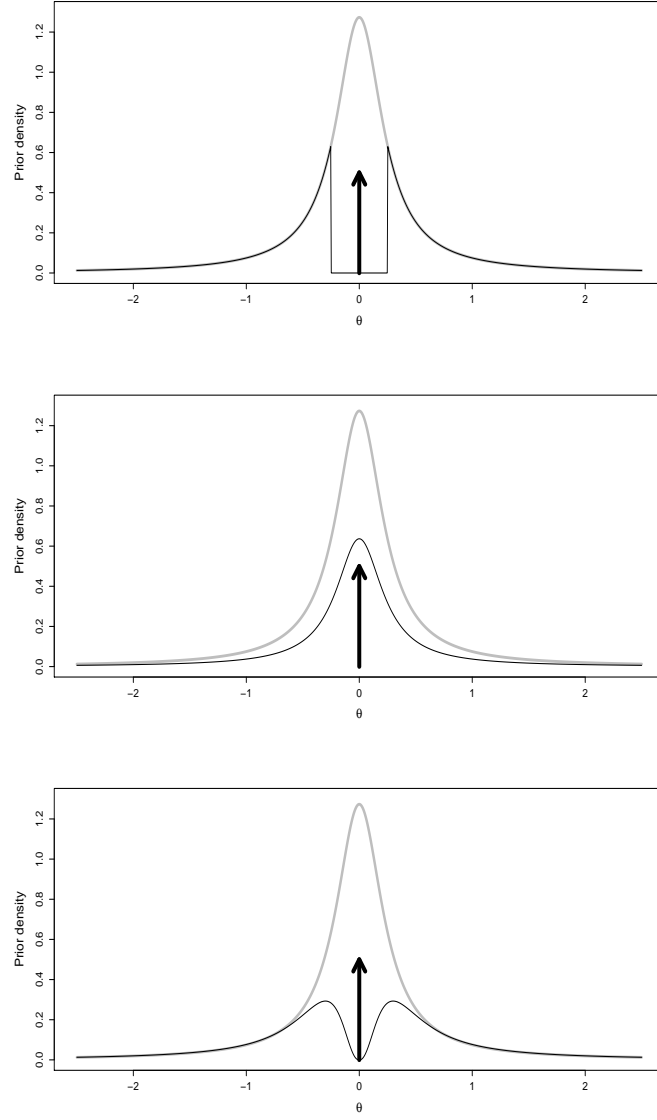


Figure 1: Marginal priors for $\theta \in \mathbb{R}$ (estimation prior $\text{Cauchy}(0, 0.0625)$ shown in grey). Top: mixture of point mass at 0 and $\text{Cauchy}(0, 0.0625)$ truncated at $\lambda = 0.25$; Middle: same with untruncated $\text{Cauchy}(0, 0.0625)$; Bottom: same as top with $\lambda \sim \text{IG}(3, 10)$

estimation prior as much as possible. This need not indicate a belief that $\theta = 0$ exactly, but could reflect that $|\theta| < \lambda$ are irrelevant, where λ is a practical significance threshold. She sets $\lambda = 0.25$ and combines a point mass at 0 with a $\text{Cauchy}(0, 0.25)$ truncated to exclude $(-0.25, 0.25)$, each with weight 0.5. The black line in Figure 1 (top) shows the resulting prior. It assigns the same $P(|\theta| > \theta_0)$ as the estimation prior for any $\theta_0 \geq 0.25$ and concentrates all probability in $(-0.25, 0.25)$ at $\theta = 0$.

The intuitive appeal of truncated priors for Bayesian tests has been argued before [Verdinelli and Wasserman, 1996, Rousseau, 2010, Klugkist and Hoijtink, 2007]. They encourage consistency between estimation and testing priors. As important limitations they require a practical significance threshold λ , and even as $n \rightarrow \infty$ there is no chance of detecting small but non-zero effect sizes. In fact, most Bayesian approaches to hypothesis testing combine point masses with untruncated priors. These include Jeffreys-Zellner-Siow priors [Jeffreys, 1961, Zellner and Siow, 1980, 1984], g and hyper-g priors [Zellner, 1986, Liang et al., 2008], unit information priors [Kass and Wasserman, 1995] and conventional priors [Bayarri and Garcia-Donato, 2007]. Although not strictly necessary, most objective Bayes approaches are also based on untruncated priors (see *e.g.* O’Hagan [1995], Berger and Pericchi [1996], Moreno et al. [1998], Berger and Pericchi [2001], Pérez and Berger [2002]). Figure 1 (middle) shows an untruncated $\text{Cauchy}(0, 0.25)$ combined with a point mass at 0. It is substantially more concentrated around 0 than the original estimation prior, *e.g.* $P(|\theta| > 0.25)$ decreased from 0.5 to 0.25. Setting a larger scale parameter for the Cauchy would not fix the issue, as the Cauchy mode would remain at 0. We view this discrepancy between estimation and testing priors as troublesome, as their underlying beliefs cannot be easily reconciled.

Suppose that the analyst goes back to the truncated Cauchy but now expresses her uncertainty in the truncation point by placing a prior $\lambda \sim G(2.5, 10)$, so that $E(\lambda) = 0.25$. Figure 1 (bottom) shows the marginal prior on θ , obtained by integrating out λ . The prior under H_1 is a smooth version of the truncated Cauchy that goes to 0 as $\theta \rightarrow 0$, hence it is a NLP (Definition 1). Relative to the estimation prior, most of the probability assigned to $\theta \approx 0$ is absorbed by the point mass, and $P(|\theta| > \theta_0)$ is roughly preserved for $\theta_0 > 0.5$. In contrast with the truncated Cauchy, it avoids setting a fixed λ and aims to detect any $\theta \neq 0$.

The example is simply meant to show a case where a NLP can be represented as a mixture of truncated distributions, and that this allows building NLPs from first principles. We formalize this intuition in the following section.

3 Non-local priors as truncation mixtures

We study the correspondence between NLPs and mixtures of truncated distributions. In Section 3.1 we show that the two classes are essentially equivalent, and in Section 3.2 we study how the nature of the mixture determines some relevant NLP characteristics. Our subsequent discussion is conditional on a given model M_k . Therefore, we simplify notation letting $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \mid M_k)$ and refer to $\dim(\boldsymbol{\theta})$ under M_k as p .

It is important to note that any NLP density under model M_k can be written as $\pi(\boldsymbol{\theta}) \propto d(\boldsymbol{\theta})\pi_u(\boldsymbol{\theta})$, where the penalty $d(\boldsymbol{\theta}) \rightarrow 0$ as $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}_0$ for any $\boldsymbol{\theta}_0 \in \Theta_k^c$ and $\pi_u(\boldsymbol{\theta})$ is an arbitrary prior. To ensure that $\pi(\boldsymbol{\theta})$ is proper we assume $\int d(\boldsymbol{\theta})\pi_u(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$. Often NLPs are directly expressed in this form (*e.g.* MOM or eMOM priors, Section 4.2), but this representation is possible even when this is not the case (*e.g.* iMOM prior, Section 4.2). Let $\pi(\boldsymbol{\theta})$ be an arbitrary NLP density, we can always write it as $\pi(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta})}{\pi_u(\boldsymbol{\theta})}\pi_u(\boldsymbol{\theta}) = d(\boldsymbol{\theta})\pi_u(\boldsymbol{\theta})$, where $\pi_u(\boldsymbol{\theta})$ is any local prior density and $d(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta})}{\pi_u(\boldsymbol{\theta})}$.

3.1 Equivalence between NLPs and truncation mixtures

We first prove that truncation mixtures define valid NLPs, and subsequently show that any NLP may be represented via truncation schemes. Given that the representation is not unique, we give two constructive approaches and discuss their relative merits.

Let $\pi_u(\boldsymbol{\theta})$ be an arbitrary prior on $\boldsymbol{\theta}$ (typically, a local prior) and let $\lambda \in \mathbb{R}^+$ be a latent truncation point. Define the conditional prior $\pi(\boldsymbol{\theta} \mid \lambda) \propto \pi_u(\boldsymbol{\theta})I(d(\boldsymbol{\theta}) > \lambda)$, where as before $d(\boldsymbol{\theta}) \rightarrow 0$ as $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}_0 \in \Theta_k^c$, and let $\pi(\lambda)$ be a marginal prior for λ . The following proposition holds.

Proposition 3.1. *Define $\pi(\boldsymbol{\theta} \mid \lambda) \propto \pi_u(\boldsymbol{\theta})I(d(\boldsymbol{\theta}) > \lambda)$, assume that $\pi(\lambda)$ places no probability mass at $\lambda = 0$ and that $\pi_u(\boldsymbol{\theta})$ is bounded in a neighbourhood around any $\boldsymbol{\theta}_0 \in \Theta_k^c$. Then $\pi(\boldsymbol{\theta}) = \int \pi(\boldsymbol{\theta} \mid \lambda)\pi(\lambda)d\lambda$ defines a non-local prior.*

Proof. See Appendix A.1. □

As a corollary to Proposition 3.1, whenever $d(\boldsymbol{\theta})$ can be expressed as the product of independent penalties $d_i(\theta_i)$ NLPs may also be induced with multiple latent truncation variables. This alternative representation can be convenient for sampling purposes (as illustrated later on) or to avoid the marginal dependency between elements in $\boldsymbol{\theta}$ induced by sharing a common truncation.

Corollary 3.1. Define $\pi_u(\boldsymbol{\theta})$ as in Proposition 3.1 and let the latent truncation points $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)' \in \mathbb{R}^{p+}$ have an absolutely continuous prior $\pi(\boldsymbol{\lambda})$. Then $\pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \propto \pi_u(\boldsymbol{\theta}) \prod_{i=1}^p I(d_i(\theta_i) > \lambda_i)$ defines a non-local prior.

Proof. Replace $I(d(\boldsymbol{\theta}) > \lambda)$ by $\prod_{i=1}^p I(d(\theta_i) > \lambda_i)$ in the proof of Proposition 3.1. Letting any λ_i go to 0 and applying the same argument delivers the result. \square

Example 3.1. Consider the linear regression model $\mathbf{y} \sim N(X\boldsymbol{\theta}, \sigma^2 I)$, where $\mathbf{y} = (y_1, \dots, y_n)'$, $\boldsymbol{\theta} \in \mathbb{R}^p$, σ^2 is known and I is the $n \times n$ identity matrix. Variable selection is conceptualized as the vanishing of any component θ_i of $\boldsymbol{\theta}$, ($i=1, \dots, p$). Accordingly, we define a NLP for $\boldsymbol{\theta}$ that penalizes $\theta_i \rightarrow 0$ with a single truncation point as in Proposition 3.1, namely $\pi(\boldsymbol{\theta} \mid \lambda) \propto N(\boldsymbol{\theta}; \mathbf{0}, \tau I) I(\prod_{i=1}^p \theta_i^2 > \lambda)$. To complete the prior specification we choose some specific form for $\pi(\lambda)$, e.g. Gamma or Inverse Gamma. Obviously, the choice of $\pi(\lambda)$ affects the properties of the marginal prior $\pi(\boldsymbol{\theta})$. We study this issue in Section 3.2. An alternative prior based on Corollary 3.1 is $\pi(\boldsymbol{\theta} \mid \lambda_1, \dots, \lambda_p) \propto N(\boldsymbol{\theta}; \mathbf{0}, \tau I) \prod_{i=1}^p I(\theta_i^2 > \lambda_i)$. This prior results in marginal independence as long as the prior on $(\lambda_1, \dots, \lambda_p)$ has independent components.

We turn attention to the complementary question: is it possible to represent any given NLP with latent truncations? The following proposition proves that such representation can always be achieved with a single truncation variable λ and an adequate choice for its prior distribution $\pi(\lambda)$.

Proposition 3.2. Let $\pi(\boldsymbol{\theta}) \propto d(\boldsymbol{\theta})\pi_u(\boldsymbol{\theta})$ be an arbitrary non-local prior and denote by $h(\lambda) = P_u(d(\boldsymbol{\theta}) > \lambda)$, where $P_u(\cdot)$ is the probability under π_u . Then $\pi(\boldsymbol{\theta})$ is the marginal prior associated to $\pi(\boldsymbol{\theta} \mid \lambda) \propto \pi_u(\boldsymbol{\theta}) I(d(\boldsymbol{\theta}) > \lambda)$ and

$$\pi(\lambda) = \frac{h(\lambda)}{E_u(d(\boldsymbol{\theta}))} \propto h(\lambda),$$

where $E_u(\cdot)$ is the expectation with respect to $\pi_u(\boldsymbol{\theta})$.

Proof. See Appendix A.2. \square

A corollary to Proposition 3.2 is that NLPs with product penalties $d(\boldsymbol{\theta}) = \prod_{i=1}^p d_i(\theta_i)$ can also be represented with multiple truncation variables.

Corollary 3.2. Let $\pi(\boldsymbol{\theta}) \propto \pi_u(\boldsymbol{\theta}) \prod_{i=1}^p d_i(\theta_i)$ be a non-local prior, denote by $h(\boldsymbol{\lambda}) = P_u(d_1(\theta_1) > \lambda_1, \dots, d_p(\theta_p) > \lambda_p)$ and assume that $h(\boldsymbol{\lambda})$ integrates to a finite constant. Then $\pi(\boldsymbol{\theta})$ is the marginal prior associated to $\pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \propto \pi_u(\boldsymbol{\theta}) \prod_{i=1}^p I(\theta_i > \lambda_i)$ and $\pi(\boldsymbol{\lambda}) \propto h(\boldsymbol{\lambda})$.

Proof. See Appendix A.3. \square

The main advantage of Corollary 3.2 is that, in spite of introducing additional latent variables, the representation greatly facilitates sampling. The technical condition that $h(\boldsymbol{\lambda})$ has a finite integral is guaranteed when $\pi_u(\boldsymbol{\theta})$ has independent components, as the result then follows by applying Proposition 3.2 to each univariate marginal. We illustrate this issue in an example where we seek to sample from the prior. Section 4 discusses the advantages for posterior sampling.

Example 3.2. Consider the Normal product MOM prior of first order $\pi(\boldsymbol{\theta}) \propto \prod_{i=1}^p \theta_i^2 N(\boldsymbol{\theta}; \mathbf{0}, \tau I)$, where τ is the prior dispersion. Setting $d(\boldsymbol{\theta}) = \prod_{i=1}^p \theta_i^2$ and $\pi_u(\boldsymbol{\theta}) = N(\boldsymbol{\theta}; \mathbf{0}, \tau I)$ in Proposition 3.2, the MOM prior can be represented using $\pi(\boldsymbol{\theta} \mid \lambda) \propto N(\boldsymbol{\theta}; \mathbf{0}, \tau I) I(\prod_{i=1}^p \theta_i^2 > \lambda)$ and the following marginal prior on λ

$$\pi(\lambda) = \frac{P(\prod_{i=1}^p \theta_i^2 / \tau > \lambda / \tau^p)}{E(\prod_{i=1}^p \theta_i^2)} = \frac{h(\lambda / \tau^p)}{\tau^p},$$

where $h(\cdot)$ is the survival function for a product of independent chi-square random variables with 1 degree of freedom (Springer and Thompson [1970] give expressions for $h(\cdot)$). Using this representation, prior draws are obtained as follows:

1. Draw $u \sim \text{Unif}(0, 1)$. Set $\lambda = P^{-1}(u)$, where $P(u) = P_\pi(\lambda \leq u)$ is the cdf associated to $\pi(\lambda)$.
2. Draw $\boldsymbol{\theta} \sim N(\mathbf{0}, \tau I) I(d(\boldsymbol{\theta}) > \lambda)$.

As important drawbacks, $P(u)$ requires Meijer G-functions and is cumbersome to evaluate for large p . Furthermore, sampling from a multivariate Normal with truncation region $\prod_{i=1}^p \theta_i^2 > \lambda$ is also non-trivial.

As an alternative representation, given the product penalty $\prod_{i=1}^p \theta_i^2$, we use Corollary 3.2 and sample from the univariate marginals. Let $P(u) = P(\lambda < u)$ be the cdf associated to $\pi(\lambda) = \frac{h(u/\tau)}{\tau}$ where $h(\cdot)$ is the survival of a χ_1^2 distribution. For $i = 1, \dots, p$ do

1. Draw $u \sim \text{Unif}(0, 1)$ and set $\lambda_i = P^{-1}(u)$.
2. Draw $\theta_i \sim N_T(0, \tau)$, where $T = \{\theta_i : \theta_i > |\lambda_i|\}$.

The function $P^{-1}(\cdot)$ can be tabulated and quickly evaluated, rendering the approach computationally efficient. Figure 2 shows 100,000 draws from univariate (left) and bivariate (right) Normal product MOM priors with $\tau = 5$.

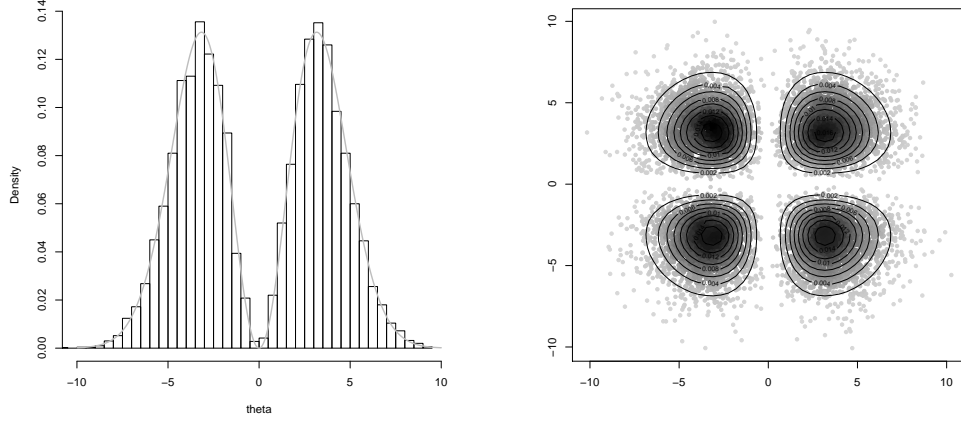


Figure 2: 10,000 independent MOM prior draws ($\tau = 5$). Lines indicate true density.

3.2 Deriving NLP properties for a given mixture

Our results so far prove a correspondence between NLPs and mixtures of truncated distributions. We now establish how the two most important characteristics of a NLP functional form, namely the penalty $d(\boldsymbol{\theta})$ and its tail behavior, depend on a given truncation scheme. It is necessary to distinguish whether a single or multiple truncation variables are used. For a common truncation variable λ , shared across $\theta_1, \dots, \theta_p$, the following holds.

Proposition 3.3. *Let $\pi(\boldsymbol{\theta})$ be the marginal non-local prior for $\pi(\boldsymbol{\theta}, \lambda) = \frac{\pi_u(\boldsymbol{\theta})}{h(\lambda)} (\prod_{i=1}^p I(d(\theta_i) > \lambda)) \pi(\lambda)$, where $h(\lambda) = P_u(d(\theta_1) > \lambda, \dots, d(\theta_p) > \lambda)$ and $\pi(\lambda)$ is absolutely continuous wrt. the Lebesgue measure on \mathbb{R}^+ . Denoting $d_{\min} = \min\{d(\theta_1), \dots, d(\theta_p)\}$, the following hold:*

1. *As $d_{\min} \rightarrow 0$, $\pi(\boldsymbol{\theta}) \propto \pi_u(\boldsymbol{\theta}) d_{\min} \pi(\lambda^*)$, where $\lambda^* \in (0, d_{\min})$. If $\pi(\lambda) \propto 1$ as $\lambda \rightarrow 0^+$ (including the case $\pi(\lambda) \propto h(\lambda)$) then $\pi(\boldsymbol{\theta}) \propto \pi_u(\boldsymbol{\theta}) d_{\min}$.*
2. *As $d_{\min} \rightarrow \infty$, $\pi(\boldsymbol{\theta})$ has tails at least as thick as $\pi_u(\boldsymbol{\theta})$. In particular, if $\int \frac{\pi(\lambda)}{h(\lambda)} d\lambda < \infty$ then $\pi(\boldsymbol{\theta}) \propto \pi_u(\boldsymbol{\theta})$.*

Proof. See Appendix A.4. □

In words, the non-local penalty is given by $d_{\min}\pi(d_{\min})$, *i.e.* only depends on the smallest individual penalty $d(\theta_1), \dots, d(\theta_p)$. This property is important as the improved learning rates for NLPs depend on the form of the penalty. Proposition 3.3 also indicates that $\pi(\boldsymbol{\theta})$ inherits its tail behavior from $\pi_u(\boldsymbol{\theta})$, which can be important for parameter estimation properties and to avoid finite sample inconsistencies [Liang et al., 2008]. Corollary 3.3 extends the results to a truncation scheme with multiple truncation points.

Corollary 3.3. *Let $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)'$ be continuous positive variables and $\pi(\boldsymbol{\theta})$ be the marginal non-local prior for $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{\pi_u(\boldsymbol{\theta})}{h(\boldsymbol{\lambda})} \prod_{i=1}^p I(d_i(\theta_i) > \lambda_i)\pi(\boldsymbol{\lambda})$, where $h(\boldsymbol{\lambda}) = P_u(d_1(\theta_1) > \lambda_1, \dots, d_p(\theta_p) > \lambda_p)$. The following hold:*

1. *As $d_i(\theta_i) \rightarrow 0$ for $i = 1, \dots, p$, $\pi(\boldsymbol{\theta}) \propto \pi_u(\boldsymbol{\theta}) \prod_{i=1}^p d_i(\theta_i)\pi(\lambda_i^*)$, where $\pi(\lambda_i^*)$ is the marginal prior for λ_i at $\lambda_i^* \in (0, d(\theta_i))$.*
2. *As $d_i(\theta_i) \rightarrow \infty$ for $i = 1, \dots, p$, $\pi(\boldsymbol{\theta})$ has tails at least as thick as $\pi_u(\boldsymbol{\theta})$. In particular, if $E(h(\boldsymbol{\lambda})^{-1}) < \infty$ under the prior on $\boldsymbol{\lambda}$, then $\pi(\boldsymbol{\theta}) \propto \pi_u(\boldsymbol{\theta})$.*

Proof. See Appendix A.5. □

In words, multiple truncation variables induce a multiplicative non-local penalty. This implies that assigning $\lambda_i \sim \pi(\lambda)$ for $i = 1, \dots, p$ induces stronger parsimony than a common $\lambda \sim \pi(\lambda)$. As an example, with multiple λ_i it may suffice that each $d_i(\lambda_i)\pi(\lambda_i)$ induces a quadratic penalty, but with a single λ an exponential penalty might be preferable.

4 Posterior sampling

We use the latent truncation characterization to derive posterior sampling algorithms, and show how setting the truncation mixture as in Proposition 3.2 and Corollary 3.2 leads to convenient simplifications. Section 4.1 provides two generic Gibbs algorithms to sample from arbitrary posteriors, and Section 4.2 adapts the algorithm to linear models under product MOM, iMOM and eMOM priors. As before, sampling is implicitly conditional on a given model M_k and we drop the conditioning on M_k to keep notation simple.

4.1 General algorithm

First consider a NLP defined by a single latent truncation variable, *i.e.* $\pi(\boldsymbol{\theta} \mid \lambda) = \pi_u(\boldsymbol{\theta})\mathbb{I}(d(\boldsymbol{\theta}) > \lambda)/h(\lambda)$, where $h(\lambda) = P_u(d(\boldsymbol{\theta}) > \lambda)$ and $\pi(\lambda)$ is an arbitrary prior on $\lambda \in \mathbb{R}^+$. The joint posterior can be expressed as

$$\pi(\boldsymbol{\theta}, \lambda \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\theta}) \frac{\pi_u(\boldsymbol{\theta})\mathbb{I}(d(\boldsymbol{\theta}) > \lambda)}{h(\lambda)} \pi(\lambda). \quad (1)$$

Sampling from $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ directly is challenging as it is usually highly multimodal. However, straightforward algebra gives the following k^{th} Gibbs iteration to sample from $\pi(\boldsymbol{\theta}, \lambda \mid \mathbf{y})$.

Algorithm 1. Gibbs sampling with a single truncation

1. Draw $\lambda^{(k)} \sim \pi(\lambda \mid \mathbf{y}, \boldsymbol{\theta}^{(k-1)}) \propto \mathbb{I}(d(\boldsymbol{\theta}) > \lambda)\pi(\lambda)/h(\lambda)$. When $\pi(\lambda) \propto h(\lambda)$ as in Proposition 3.2, $\lambda^{(k)} \sim \text{Unif}(0, d(\boldsymbol{\theta}^{(k-1)}))$.
2. Draw $\boldsymbol{\theta}^{(k)} \sim \pi(\boldsymbol{\theta} \mid \mathbf{y}, \lambda^{(k)}) \propto \pi_u(\boldsymbol{\theta})\mathbb{I}(d(\boldsymbol{\theta}) > \lambda^{(k)})$.

That is, $\lambda^{(k)}$ is sampled from a univariate distribution that reduces to a uniform when setting $\pi(\lambda) \propto h(\lambda)$, and $\boldsymbol{\theta}^{(k)}$ from a truncated version of $\pi_u(\cdot)$. For instance, $\pi_u(\cdot)$ may be an estimation prior that allows easy posterior sampling. As a difficulty, the truncation region $\{\boldsymbol{\theta} : d(\boldsymbol{\theta}) > \lambda^{(k)}\}$ is non-linear and non-convex so that jointly sampling $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ may be challenging. One may apply a Gibbs step to each element in $\theta_1, \dots, \theta_p$ sequentially, which only requires univariate truncated draws from $\pi_u(\cdot)$, but the mixing of the chain may suffer.

The multiple truncation representation in Corollary 3.2 provides a convenient alternative algorithm. Consider $\pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) = \pi_u(\boldsymbol{\theta}) \prod_{i=1}^p \mathbb{I}(d_i(\theta_i) > \lambda_i)\pi(\boldsymbol{\lambda})/h(\boldsymbol{\lambda})$, where $h(\boldsymbol{\lambda}) = P_u(d_1(\theta_1) > \lambda_1, \dots, d_p(\theta_p) > \lambda_p)$ and $\pi(\boldsymbol{\lambda})$ is an arbitrary prior. The following steps define the k^{th} iteration in a Gibbs sampling scheme:

Algorithm 2. Gibbs sampling with multiple truncations

1. Draw $\boldsymbol{\lambda}^{(k)} \sim \pi(\boldsymbol{\lambda} \mid \mathbf{y}, \boldsymbol{\theta}^{(k-1)}) = \prod_{i=1}^p \text{Unif}(\lambda_i; 0, d_i(\theta_i)) \frac{\pi(\boldsymbol{\lambda})}{h(\boldsymbol{\lambda})}$. If $\pi(\boldsymbol{\lambda}) \propto h(\boldsymbol{\lambda})$ as in Corollary 3.2, $\lambda_i^{(k)} \sim \text{Unif}(0, d_i(\theta_i))$.
2. Draw $\boldsymbol{\theta}^{(k)} \sim \pi(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\lambda}^{(k)}) \propto \pi_u(\boldsymbol{\theta}) \prod_{i=1}^p \mathbb{I}(d_i(\theta_i) > \lambda_i^{(k)})$

Now the truncation region in Step 2 is defined by hyper-rectangles, which facilitates sampling. As in Algorithm 1, by setting the prior conveniently Step 1 avoids evaluating $\pi(\boldsymbol{\lambda})$ and $h(\boldsymbol{\lambda})$.

4.2 Linear models

We adapt Algorithm 2 to a Normal linear regression $\mathbf{y} \sim N(X\boldsymbol{\theta}, \phi I)$ with unknown variance ϕ and a product NLP on $\boldsymbol{\theta}$. We consider the three following priors

$$\pi_M(\boldsymbol{\theta} \mid \phi) = \prod_{i=1}^p \frac{\theta_i^2}{\tau\phi} N(\boldsymbol{\theta}; \mathbf{0}; \tau\phi I) \quad (2)$$

$$\pi_I(\boldsymbol{\theta} \mid \phi) = \prod_{i=1}^p \frac{(\tau\phi)^{\frac{1}{2}}}{\sqrt{\pi}\theta_i^2} \exp\left\{-\frac{\tau\phi}{\theta_i^2}\right\} \quad (3)$$

$$\pi_E(\boldsymbol{\theta} \mid \phi) = \prod_{i=1}^p \exp\left\{\sqrt{2} - \frac{\tau\phi}{\theta_i^2}\right\} N(\boldsymbol{\theta}; \mathbf{0}; \tau\phi I), \quad (4)$$

where π_M , π_I and π_E are the product MOM, iMOM and eMOM priors (respectively). As in Johnson and Rossell [2012], we set the prior $\phi \sim \text{IG}(a_\phi/2, b_\phi/2)$ and let τ be a user-specified prior dispersion. To set a hyperprior on τ or determine it in a data-adaptively manner see Rossell et al. [2012], and for objective Bayes alternatives see *e.g.* Consonni and La Rocca [2010].

For all three priors, Step 2 in Algorithm 2 samples from a multivariate Normal with rectangular truncation around $\mathbf{0}$, for which we developed an efficient algorithm. Kotecha and Djuric [1999] and Rodriguez-Yam et al. [2004] proposed Gibbs after orthogonalization strategies that result in low serial correlation, and Wilhelm and Manjunath [2010] implemented the approach in the R package `tmvtnorm` under restrictions of the type $l \leq \theta_i \leq u$. Here we require sampling under $d_i(\theta_i) \geq l$, which defines a non-convex region. Our adapted algorithm is in Appendix A.6 and implemented in our R package `mombf`. An important property is that the algorithm produces independent samples when the posterior probability of the truncation region becomes negligible. Since NLPs only assign high posterior probability to a model when the posterior for non-zero coefficients is well shifted from the origin, the truncation region is indeed often negligible. We outline the full algorithm separately for each prior.

4.2.1 Product MOM prior

Straightforward algebra shows that the full conditional posteriors are

$$\begin{aligned}\pi(\boldsymbol{\theta} \mid \phi, \mathbf{y}) &\propto \left(\prod_{i=1}^p \theta_i^2 \right) N(\boldsymbol{\theta}; \mathbf{m}, \phi S^{-1}) \\ \pi(\phi \mid \boldsymbol{\theta}, \mathbf{y}) &= \text{IG} \left(\frac{a_\phi + n + 3p}{2}, \frac{b_\phi + s_R^2 + \boldsymbol{\theta}'\boldsymbol{\theta}/\tau}{2} \right),\end{aligned}\quad (5)$$

where $S = X'X + \tau^{-1}I$, $\mathbf{m} = S^{-1}X'\mathbf{y}$ and $s_R^2 = (\mathbf{y} - X\boldsymbol{\theta})'(\mathbf{y} - X\boldsymbol{\theta})$ is the sum of squared residuals. Corollary 3.2 allows to represent the product MOM prior as

$$\pi(\boldsymbol{\theta} \mid \phi, \boldsymbol{\lambda}) = N(\boldsymbol{\theta}; \mathbf{0}, \tau\phi I) \prod_{i=1}^p \text{I} \left(\frac{\theta_i^2}{\tau\phi} > \lambda_i \right) \frac{1}{h(\lambda_i)} \quad (6)$$

marginalized with respect to $\pi(\lambda_i) = h(\lambda_i) = P \left(\frac{\theta_i^2}{\tau\phi} > \lambda_i \mid \phi \right)$, where $h(\cdot)$ is the survival of a chi-square with 1 degree of freedom. Algorithm 2 and straightforward algebra give the k^{th} Gibbs iteration as

1. $\phi^{(k)} \sim \text{IG} \left(\frac{a_\phi + n + 3p}{2}, \frac{b_\phi + s_R^2 + (\boldsymbol{\theta}^{(k-1)})'\boldsymbol{\theta}^{(k-1)}/\tau}{2} \right)$
2. $\boldsymbol{\lambda}^{(k)} \sim \pi(\boldsymbol{\lambda} \mid \boldsymbol{\theta}^{(k-1)}, \phi^{(k)}, \mathbf{y}) = \prod_{i=1}^p \text{I} \left(\frac{(\theta_i^{(k-1)})^2}{\tau\phi^{(k)}} > \lambda_i \right)$
3. $\boldsymbol{\theta}^{(k)} \sim \pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}^{(k)}, \phi^{(k)}, \mathbf{y}) = N(\boldsymbol{\theta}; \mathbf{m}, \phi^{(k)} S^{-1}) \prod_{i=1}^p \text{I} \left(\frac{\theta_i^2}{\tau\phi^{(k)}} > \lambda_i \right).$

Step 1 samples unconditionally on $\boldsymbol{\lambda}$, so that no efficiency is lost for introducing these latent variables. Step 2 and 3 jointly sample from $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ (respectively), the latter requiring truncated multivariate Normal draws.

4.2.2 Product iMOM prior

We focus on sampling from the posterior of a model with $p \leq n$. The full conditional posteriors are

$$\begin{aligned}\pi(\boldsymbol{\theta} \mid \phi, \mathbf{y}) &\propto \left(\prod_{i=1}^p \frac{\sqrt{\tau\phi}}{\theta_i^2} e^{-\frac{\tau\phi}{\theta_i^2}} \right) N(\boldsymbol{\theta}; \mathbf{m}, \phi S^{-1}) \\ \pi(\phi \mid \boldsymbol{\theta}, \mathbf{y}) &= e^{-\tau\phi \sum_{i=1}^p \theta_i^{-2}} \text{IG} \left(\phi; \frac{a_\phi + n - p}{2}, \frac{b_\phi + s_R^2}{2} \right),\end{aligned}\quad (7)$$

where $S = X'X$, $\mathbf{m} = S^{-1}X'\mathbf{y}$ and $s_R^2 = (\mathbf{y} - X\boldsymbol{\theta})'(\mathbf{y} - X\boldsymbol{\theta})$.

Now, the iMOM prior in (4) can be written as $\pi_I(\boldsymbol{\theta} \mid \phi) =$

$$N(\boldsymbol{\theta}; \mathbf{0}; \tau_N \phi \mathbf{I}) \prod_{i=1}^p \frac{\frac{\sqrt{\tau\phi}}{\sqrt{\pi\theta_i^2}} e^{-\frac{\phi\tau}{\theta_i^2}}}{N(\theta_i; 0, \tau_N \phi)} = N(\boldsymbol{\theta}; \mathbf{0}; \tau_N \phi \mathbf{I}) \prod_{i=1}^p d_i(\theta_i, \phi). \quad (8)$$

While in principle any value of τ_N may be used, setting $\tau_N \geq 2\tau$ guarantees $d(\theta_i, \phi)$ to be monotone increasing in θ_i^2 , so that its inverse exists (Appendix A.7). By default we set $\tau_N = 2\tau$. Following Corollary 3.2 we represent (8) using latent variables $\boldsymbol{\lambda}$, *i.e.*

$$\pi(\boldsymbol{\theta} \mid \phi, \boldsymbol{\lambda}) = N(\boldsymbol{\theta}; \mathbf{0}, \tau_N \phi \mathbf{I}) \prod_{i=1}^p \mathbf{I}(d(\theta_i, \phi) > \lambda_i) \frac{1}{h(\lambda_i)} \quad (9)$$

and $\pi(\boldsymbol{\lambda}) = \prod_{i=1}^p h(\lambda_i)$, where $h(\lambda_i) = P(d(\theta_i, \phi) > \lambda_i)$ which we need not evaluate. Algorithm 2 gives the following MH within Gibbs procedure.

1. MH step

(a) Propose $\phi^* \sim \text{IG}\left(\phi; \frac{a_\phi + n - p}{2}, \frac{b_\phi + s_R^2}{2}\right)$

(b) Set $\phi^{(k)} = \phi^*$ with probability $\min\left\{1, e^{(\phi^{(k-1)} - \phi^*)\tau \sum_{i=1}^p \theta_i^{-2}}\right\}$, else $\phi^{(k)} = \phi^{(k-1)}$.

2. $\boldsymbol{\lambda}^{(k)} \sim \prod_{i=1}^p \text{Unif}\left(\lambda_i; 0, d(\theta_i^{(k-1)}, \phi^{(k)})\right)$

3. $\boldsymbol{\theta}^{(k)} \sim N(\boldsymbol{\theta}; \mathbf{m}, \phi^{(k)} S^{-1}) \prod_{i=1}^p \mathbf{I}\left(d(\theta_i, \phi^{(k)}) > \lambda_i^{(k)}\right)$.

Step 3 requires the inverse $d^{-1}(\cdot)$, which can be evaluated efficiently combining an asymptotic approximation with a linear interpolation search (Appendix A.7). As a token, 10,000 draws for $p = 2$ variables required 0.58 seconds on a 2.8 GHz processor running OS X 10.6.8.

4.2.3 Product eMOM prior

The full conditional posteriors are

$$\begin{aligned} \pi(\boldsymbol{\theta} \mid \phi, \mathbf{y}) &\propto \left(\prod_{i=1}^p e^{-\frac{\tau\phi}{\theta_i^2}} \right) N(\boldsymbol{\theta}; \mathbf{m}, \phi S^{-1}) \\ \pi(\phi \mid \boldsymbol{\theta}, \mathbf{y}) &\propto e^{-\sum_{i=1}^p \frac{\tau\phi}{\theta_i^2}} \text{IG}\left(\phi; \frac{a^*}{2}, \frac{b^*}{2}\right), \end{aligned} \quad (10)$$

where $S = X'X + \tau^{-1}I$, $\mathbf{m} = S^{-1}X'\mathbf{y}$, $a^* = a_\phi + n + p$, $b^* = b_\phi + s_R^2 + \boldsymbol{\theta}'\boldsymbol{\theta}/\tau$ and $s_R^2 = (\mathbf{y} - X\boldsymbol{\theta})'(\mathbf{y} - X\boldsymbol{\theta})$. Corollary 3.2 represents the product eMOM prior $\pi_E(\boldsymbol{\theta})$ in (4) as

$$\pi(\boldsymbol{\theta} \mid \phi, \boldsymbol{\lambda}) = N(\boldsymbol{\theta}; \mathbf{0}, \tau\phi I) \prod_{i=1}^p \mathbf{I}\left(e^{\sqrt{2} - \frac{\tau\phi}{\theta_i^2}} > \lambda_i\right) \frac{1}{h(\lambda_i)} \quad (11)$$

marginalized with respect to $\pi(\lambda_i) = h(\lambda_i) = P\left(e^{\sqrt{2} - \frac{\tau\phi}{\theta_i^2}} > \lambda_i \mid \phi\right)$. Again $h(\lambda_i)$ has no simple form but is not required by Algorithm 2, which gives the k^{th} Gibbs iteration

1. $\phi^{(k)} \sim e^{-\sum_{i=1}^p \frac{\tau\phi}{\theta_i^2}} \text{IG}\left(\phi; \frac{a^*}{2}, \frac{b^*}{2}\right)$
 - (a) Propose $\phi^* \sim \text{IG}\left(\phi; \frac{a^*}{2}, \frac{b^*}{2}\right)$
 - (b) Set $\phi^{(k)} = \phi^*$ with probability $\min\left\{1, e^{(\phi^{(k-1)} - \phi^*)\tau \sum_{i=1}^p \theta_i^{-2}}\right\}$, else $\phi^{(k)} = \phi^{(k-1)}$.
2. $\boldsymbol{\lambda}^{(k)} \sim \prod_{i=1}^p \text{Unif}\left(\lambda_i; 0, e^{\sqrt{2} - \frac{\tau\phi}{\theta_i^2}}\right)$
3. $\boldsymbol{\theta}^{(k)} \sim N(\boldsymbol{\theta}; \mathbf{m}, \phi^{(k)} S^{-1}) \prod_{i=1}^p \mathbf{I}\left(\theta_i^2 > \left\lceil \frac{\phi\tau}{\log(\lambda_i) - \sqrt{2}} \right\rceil\right)$.

5 Examples

We assess the performance of the posterior sampling algorithms proposed in Section 4, as well as the use of NLPs for high-dimensional parameter estimation. Section 5.1 shows a bivariate example designed to illustrate the main posterior sampling issues. Section 5.2 studies a high-dimensional case where $p = n$ and compares the model-averaging estimators induced by NLPs with benchmark [Fernández et al., 2001] and hyper-g priors [Liang et al., 2008], SCAD [Fan and Li, 2001] and LASSO [Tibshirani, 1996]. In all examples we use the default prior dispersions $\tau = 0.358, 0.133, 0.119$ for product MOM, iMOM and eMOM priors (respectively), all of which assign 0.01 prior probability to $|\theta_i/\sqrt{\phi}| < 0.2$ [Johnson and Rossell, 2010], and $\phi \sim \text{IG}(0.01/2, 0.01/2)$.

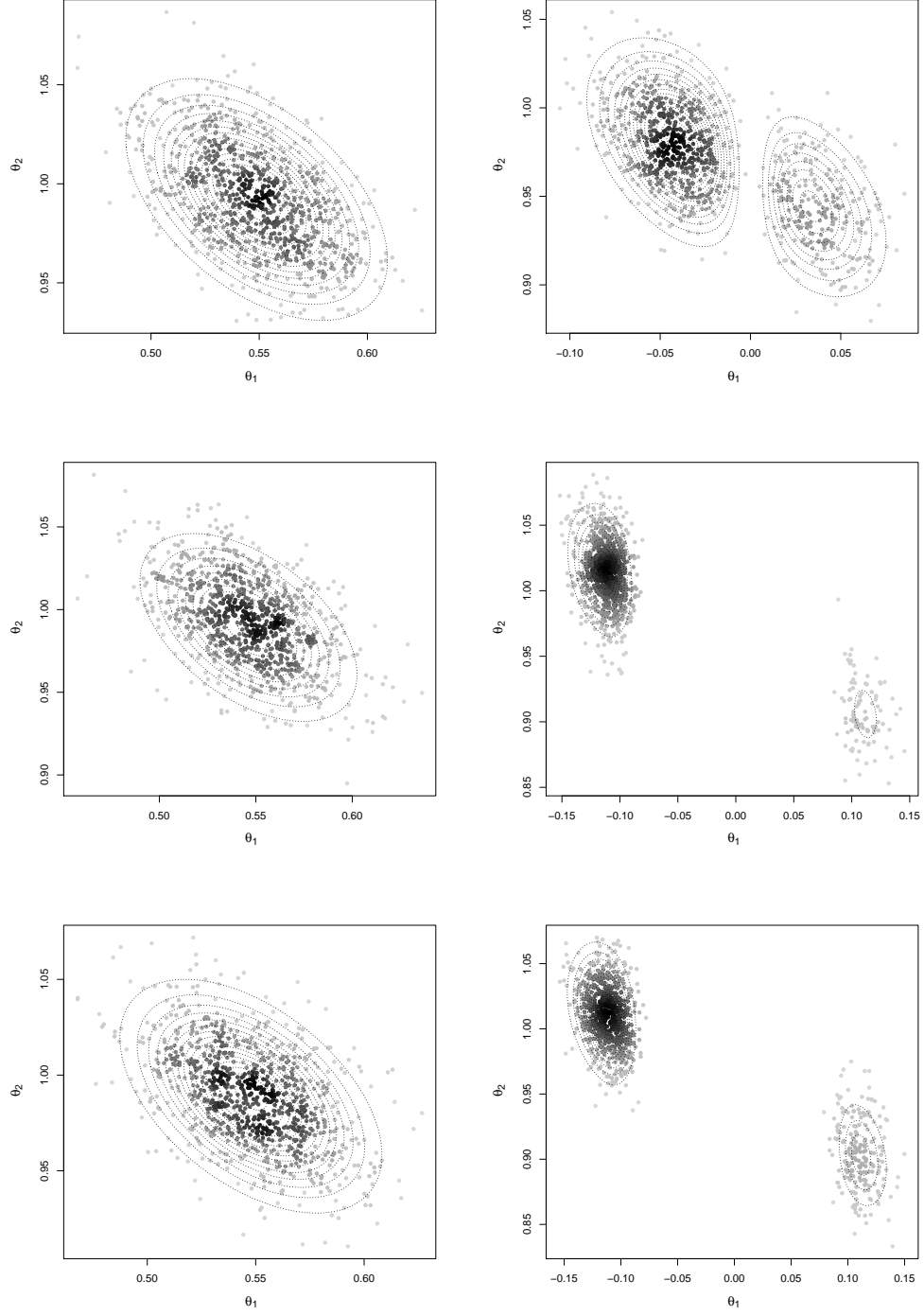


Figure 3: 900 Gibbs draws after a 100 burn-in when $\theta = (0.5, 1)'$ (left) and $\theta = (0, 1)'$ (right). Contours indicate posterior density. Top: MOM ($\tau = 0.358$); Middle: iMOM ($\tau = 0.133$); Bottom: eMOM ($\tau = 0.119$)

$\theta_1 = 0.5, \theta_2 = 1$			
	MOM	iMOM	eMOM
$\theta_1 = 0, \theta_2 = 0$	0	0	0
$\theta_1 = 0, \theta_2 \neq 0$	2.8e-78	2.72e-78	6.86e-79
$\theta_1 \neq 0, \theta_2 = 0$	1.95e-191	3.82-e191	5.90e-191
$\theta_1 \neq 0, \theta_2 \neq 0$	1	1	1

$\theta_1 = 0, \theta_2 = 1$			
	MOM	iMOM	eMOM
$\theta_1 = 0, \theta_2 = 0$	1.69e-225	4.39e-225	1.08e-224
$\theta_1 = 0, \theta_2 \neq 0$	0.999	1	1
$\theta_1 \neq 0, \theta_2 = 0$	1.82e-193	1.64e-192	6.80e-192
$\theta_1 \neq 0, \theta_2 \neq 0$	8.83e-05	3.30e-09	3.17e-09

Table 1: Posterior probabilities

$\theta_1 = 0.5, \theta_2 = 1$			
	MOM	iMOM	eMOM
θ_1	0.096	0.110	0.018
θ_2	0.034	0.134	0.019
ϕ	-0.016	0.069	0.027

$\theta_1 = 0, \theta_2 = 1$			
	MOM	iMOM	eMOM
θ_1	0.115	0.032	0.049
θ_2	0.134	0.122	0.042
ϕ	-0.040	0.327	0.353

Table 2: Serial correlation in Gibbs sampling algorithm

5.1 Posterior samples for a given model

We simulate 1,000 realizations from $y_i \sim N(\theta_1 x_{1i} + \theta_2 x_{2i}, 1)$, where (x_{1i}, x_{2i}) are drawn from a bivariate Normal with $E(x_{1i}) = E(x_{2i}) = 0$, $V(x_{1i}) = V(x_{2i}) = 2$, $\text{Cov}(x_{1i}, x_{2i}) = 1$.

We first consider a scenario where both variables have non-zero coefficients $\theta_1 = 0.5$, $\theta_2 = 1$, and compute posterior probabilities for the four possible models. We assign equal a priori probabilities and obtain exact integrated likelihoods using functions `pmomMarginalU`, `pimomMarginalU` and `pemomMarginalU` in the `mombf` package (the former is available in closed-form, for the latter two we used 10^6 importance samples). The posterior probability assigned to the full model under all three priors is 1 (up to rounding) (Table 1). Figure 3 (left) shows 900 Gibbs draws (100 burn-in) obtained under the full model. The posterior mass is well-shifted away from 0 and resembles an elliptical shape for the three priors, as expected. Table 2 gives the first-order auto-correlations, which are very small. This example reflects the advantages of the orthogonalization strategy, which is particularly efficient as the latent truncation becomes negligible.

We now set $\theta_1 = 0$, $\theta_2 = 1$ and keep $n = 1000$ and (x_{1i}, x_{2i}) as before. We simulated several data sets and in most cases did not observe a noticeable multi-modality in the posterior density. We portray a specific simulation that did exhibit multi-modality, as this poses a greater challenge from a sampling perspective. Table 1 shows that the data-generating model adequately concentrated the posterior mass. Although the full model was clearly dismissed in light of the data, as an exercise we drew from its posterior. Figure 3 (right) shows 900 Gibbs draws after a 100 burn-in, and Table 2 indicates the auto-correlation. The sampled values adequately captured the multi-modal, non-elliptical posterior.

5.2 High-dimensional estimation

We consider full model averaging on θ and reproduce the high-dimensional simulation study in Johnson and Rossell [2012]. We set $p = n = 50, 100, 200$, $\theta_i = 0$ for $i = 1, \dots, p-5$, the remaining 5 coefficients to $(0.6, 1.2, 1.8, 2.4, 3)$ and residual variance $\phi = 1, 4$. We simulated 1,000 data sets under each setup. Covariate values are sampled from $\mathbf{x} \sim N(\mathbf{0}, \Sigma)$, where Σ is either diagonal or a matrix with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.25$ for $i \neq j$. We remark that these are population correlations, the 95% quantiles of the absolute sample correlations when $\Sigma_{ij} = 0$ were 0.28, 0.20 and 0.14 for $n = 50, 100, 200$ (respectively), and 0.45, 0.39, 0.35 when $\Sigma_{ij} = 0.25$.

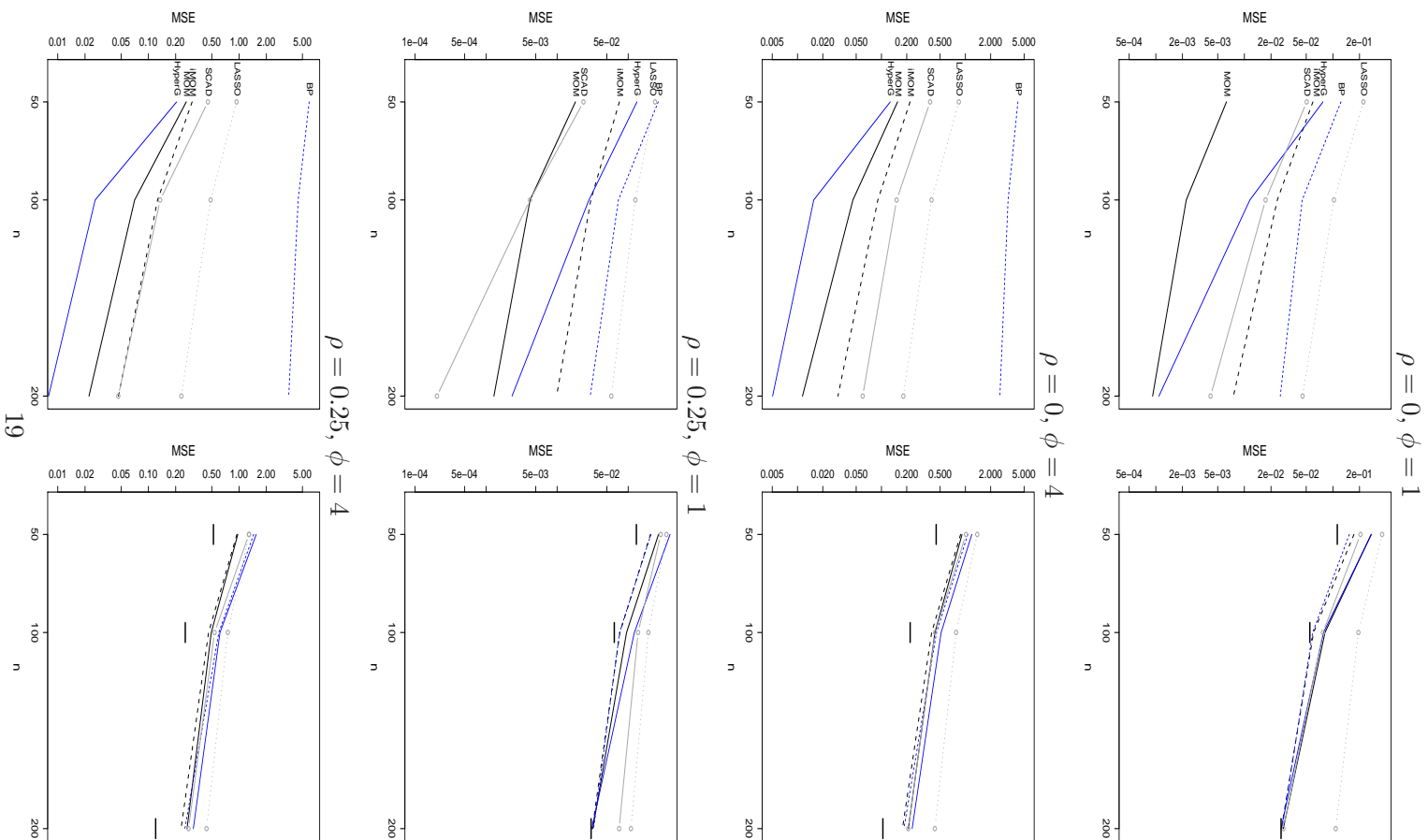


Figure 4: Average MSE for $\theta_i = 0$ (left) and $\theta_i \neq 0$ (right)

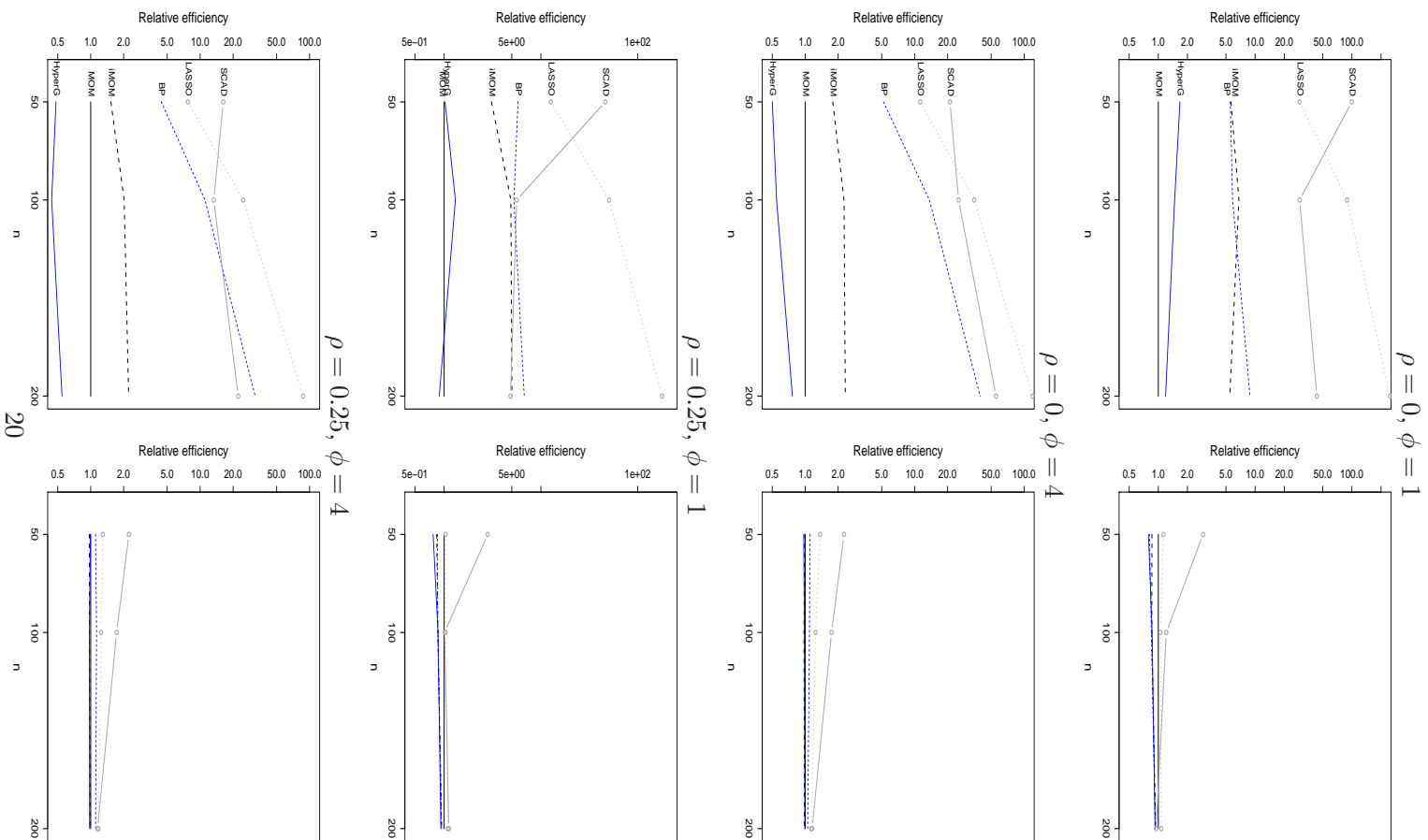


Figure 5: Mean CI width for $\theta_i = 0$ (left) and $\theta_i \neq 0$ (right)

n	MOM	iMOM	BM	HG	SCAD	LASSO
$\rho = 0, \phi = 1$						
50	1.000	1.000	0.998	0.997	1.000	1.000
100	1.000	1.000	0.999	1.000	1.000	0.999
200	1.000	1.000	1.000	1.000	1.000	0.999
$\rho = 0, \phi = 4$						
50	1.000	1.000	0.939	1.000	1.000	1.000
100	1.000	1.000	0.945	1.000	1.000	0.999
200	1.000	1.000	0.952	1.000	1.000	0.999
$\rho = 0.25, \phi = 1$						
50	1.000	1.000	0.997	0.995	1.000	1.000
100	1.000	1.000	0.999	0.999	1.000	0.999
200	1.000	1.000	1.000	1.000	1.000	0.999
$\rho = 0.25, \phi = 4$						
50	1.000	1.000	0.933	1.000	1.000	1.000
100	1.000	1.000	0.941	1.000	1.000	0.999
200	1.000	1.000	0.951	1.000	1.000	0.999

Table 3: Frequentist coverage of 95% CI intervals for zero coefficients

Let δ be the model indicator. For each simulated data set, we drew 5,000 samples $\delta_1, \dots, \delta_{5000}$ from $P(\delta \mid \mathbf{y})$ (500 burn-in) with the Gibbs scheme used by Johnson and Rossell [2012]. We estimated model probabilities from the number of visits to each model, and obtained posterior draws for $\boldsymbol{\theta}$ using the algorithms in Section 4.2. For benchmark (BP) and hyper-g (HG) priors we used the R package BMA [Raftery et al., 2013]. We computed the mean and empirical 2.5% and 97.5% quantiles from the obtained samples.

We report estimation Mean Square Errors (MSE), average posterior interval widths and their empirical coverage probabilities. For comparison, we used SCAD and LASSO to estimate regression coefficients (penalty parameter set via 10-fold cross-validation) and used Bootstrap (1,000 samples) to obtain 95% confidence intervals. Figure 4 shows the MSE separately across all $\theta_i = 0$ (left) and $\theta_i \neq 0$ (right). Strong differences between methods were observed for zero coefficients, whereas for non-zero coefficients most methods performed similarly. When $\theta_i = 0$, MOM had lowest or second lowest MSE in all scenarios, with either SCAD or HG as best performing alternatives. iMOM had higher MSE than MOM, although the differences were smaller when $\phi = 4$. Interestingly, for $\theta_i \neq 0$ iMOM showed slightly lower MSE than other methods in almost all situations. In fact, when $\phi = 1$ its MSE

n	MOM	iMOM	BM	HG	SCAD	LASSO
$\rho = 0, \phi = 1$						
50	0.940	0.941	0.943	0.890	0.981	0.813
100	0.960	0.941	0.944	0.936	0.952	0.774
200	0.957	0.952	0.947	0.953	0.938	0.767
$\rho = 0, \phi = 4$						
50	0.836	0.837	0.823	0.783	0.953	0.824
100	0.870	0.872	0.863	0.840	0.958	0.793
200	0.914	0.919	0.902	0.906	0.959	0.786
$\rho = 0.25, \phi = 1$						
50	0.954	0.939	0.945	0.862	0.982	0.837
100	0.966	0.943	0.945	0.925	0.927	0.781
200	0.960	0.950	0.946	0.952	0.929	0.756
$\rho = 0.25, \phi = 4$						
50	0.847	0.841	0.802	0.763	0.958	0.854
100	0.864	0.868	0.853	0.822	0.945	0.807
200	0.899	0.905	0.889	0.873	0.941	0.793

Table 4: Frequentist coverage of 95% CI intervals for non-zero coefficients

was very close to that of the least squares oracle estimator by $n = p = 100$ (Figure 4, right panel, black horizontal segments).

Figure 5 shows the 95% interval average width of each method relative to that of MOM. Again methods differ most for zero coefficients, where MOM and HG-based intervals are between 5 and 100 fold shorter than for other methods. The coverage probability of these intervals achieved the desired 0.95 for all approaches (Table 3), with a slight under-coverage for BP. However, for non-zero coefficients significant under-coverages were observed. MOM, iMOM and BM had roughly 0.95 coverage when $\phi = 1$, and an under-coverage when $\phi = 4$ that improved with larger n . HG priors had a substantial under-coverage when $n = 50$ in all scenarios, although it also improved with n . The undercoverage of Bayesian methods when $\phi = 4$ (*i.e.* smaller signal-to-noise ratio) suggests a certain over-shrinkage in finite sample sizes. SCAD was best in preserving the target 0.95 coverage, whereas LASSO exhibited a marked under-coverage that did not improve with sample size.

6 Discussion

We studied the use of NLPs for parameter estimation, with an emphasis on high-dimensional regression. The work is founded on representing NLPs as mixtures of truncated distributions. The representation is always possible, allows defining NLPs starting from arbitrary estimation priors and greatly facilitates posterior sampling. We provided sampling algorithms for three NLP families (MOM, iMOM and eMOM) in a linear regression setup. Beyond their computational appeal, latent truncations build NLPs from intuitive first principles.

We observed promising results. On one hand, posterior samples exhibited low serial correlation and captured multi-modalities. Additionally, the model averaging posterior mean induced by NLPs performed remarkably in our high-dimensional simulations. In particular, MOM was the best overall performer, combining small MSE and interval widths for zero coefficients with competitive MSE and coverage for non-zero coefficients. These findings suggest that NLPs may prove particularly useful in ultra-high dimensional sparse setups. Johnson and Rossell [2012] showed that NLPs provided advantages to select the data-generating model in a similar setup. Hence, our results show that it is not only possible to use the same prior for estimation and selection, but may indeed be desirable. We remark that we used default informative priors, which are relatively popular for testing, but perhaps less readily adopted for estimation. In fact, although hyper-g priors were originally designed for model selection, they also performed fairly well for estimation.

The proposed construction of NLPs enables feasible estimation and model determination for problems beyond linear regression. In particular, inference based on MCMC in several model classes may be easily achieved by applying our results to extend available posterior simulation strategies. Promising applications include, but are not limited to, generalized linear, graphical and mixture models.

A Proofs and Miscellanea

A.1 Proof of Proposition 3.1

The goal is to show that for all $\epsilon > 0$ there exists $\eta > 0$ such that $d(\boldsymbol{\theta}) < \eta$ implies $\pi(\boldsymbol{\theta}) < \epsilon$. By construction, the conditional prior density is $\pi(\boldsymbol{\theta} \mid \lambda) = \pi_u(\boldsymbol{\theta})\mathbf{I}(d(\boldsymbol{\theta}) > \lambda)/h(\lambda)$, where $h(\lambda) = P_u(d(\boldsymbol{\theta}) > \lambda) = \int \pi_u(\boldsymbol{\theta})\mathbf{I}(d(\boldsymbol{\theta}) >$

$\lambda)d\boldsymbol{\theta}$. Let $\boldsymbol{\theta}$ be a value such that $d(\boldsymbol{\theta}) < \eta$, and express the prior density as

$$\begin{aligned} \pi(\boldsymbol{\theta}) &= \int \pi(\boldsymbol{\theta} \mid \lambda) \pi(\lambda) d\lambda = \\ &= \int_{\lambda \leq \eta} \frac{\pi_u(\boldsymbol{\theta}) \mathbb{I}(d(\boldsymbol{\theta}) > \lambda)}{h(\lambda)} \pi(\lambda) d\lambda + \int_{\lambda > \eta} \frac{\pi_u(\boldsymbol{\theta}) \mathbb{I}(d(\boldsymbol{\theta}) > \lambda)}{h(\lambda)} \pi(\lambda) d\lambda \end{aligned} \quad (12)$$

The second term in (12) is 0, as by assumption $d(\boldsymbol{\theta}) < \eta$. Now, consider that for $\lambda \leq \eta$, $h(\lambda) = P_u(d(\boldsymbol{\theta}) > \lambda)$ is minimized at $\lambda = \eta$, and therefore (12) can be bounded by

$$\pi(\boldsymbol{\theta}) \leq \frac{\pi_u(\boldsymbol{\theta}) \int_{\lambda \leq \eta} \mathbb{I}(d(\boldsymbol{\theta}) > \lambda) \pi(\lambda) d\lambda}{h(\eta)} = \frac{\pi_u(\boldsymbol{\theta}) P(\lambda < \min\{\eta, d(\boldsymbol{\theta})\})}{h(\eta)} \quad (13)$$

Notice that the numerator can be made arbitrarily small by decreasing η , since $\pi_u(\boldsymbol{\theta})$ is bounded around $\boldsymbol{\theta}_0$, by assumption there is no prior mass at $\lambda = 0$ so that the cdf in the numerator converges to 0 as $\eta \rightarrow 0$, and that denominator converges to 1 as $\eta \rightarrow 0$. That is, it is possible to choose η such that $\pi(\boldsymbol{\theta}) \leq \epsilon$, which gives the result. \square

A.2 Proof of Proposition 3.2

We first note that in order for $\pi(\boldsymbol{\theta})$ to be proper the random variable $d(\boldsymbol{\theta})$ must have finite expectation with respect to $\pi_u(\boldsymbol{\theta})$. Now, the marginal prior for $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}) = \int \frac{\pi_u(\boldsymbol{\theta}) \mathbb{I}(d(\boldsymbol{\theta}) > \lambda)}{P_u(d(\boldsymbol{\theta}) > \lambda)} \pi(\lambda) d\lambda = \pi_u(\boldsymbol{\theta}) \int_0^{d(\boldsymbol{\theta})} \frac{\pi(\lambda)}{h(\lambda)} d\lambda. \quad (14)$$

Suppose we set $\pi(\lambda) \propto h(\lambda)$, which we can do as long as $\int h(\lambda) d\lambda < \infty$. Then $\pi(\boldsymbol{\theta}) \propto \pi_u(\boldsymbol{\theta}) d(\boldsymbol{\theta})$, which proves the result. The only step left is to show that indeed $\int h(\lambda) d\lambda < \infty$. In general

$$\int h(\lambda) d\lambda = \int P_u(d(\boldsymbol{\theta}) > \lambda) d\lambda = \int S_{d(\boldsymbol{\theta})}(\lambda) d\lambda, \quad (15)$$

where $S_{d(\boldsymbol{\theta})}(\lambda)$ is the survival function of the positive random variable $d(\boldsymbol{\theta})$ and therefore (15) is equal to its expectation $E_u(d(\boldsymbol{\theta}))$ with respect to $\pi_u(\boldsymbol{\theta})$, which is finite as discussed at the beginning of the proof. \square

A.3 Proof of Corollary 3.2

Analogously to the proof of Proposition 3.2 the marginal prior for $\boldsymbol{\theta}$ is $\pi(\boldsymbol{\theta}) =$

$$\begin{aligned} & \int \cdots \int \frac{\pi_u(\boldsymbol{\theta}) \prod_{i=1}^p \mathbb{I}(d_i(\theta_i) > \lambda_i)}{P_u(d_1(\theta_1) > \lambda_1, \dots, d_p(\theta_p) > \lambda_p)} \pi(\boldsymbol{\lambda}) d\lambda_1, \dots, d\lambda_p = \\ & \pi_u(\boldsymbol{\theta}) \int_0^{d_1(\theta_1)} \cdots \int_0^{d_p(\theta_p)} \frac{\pi(\boldsymbol{\lambda})}{h(\boldsymbol{\lambda})} d\lambda_1, \dots, d\lambda_p \propto \pi_u(\boldsymbol{\theta}) \prod_{i=1}^p d_i(\theta_i), \end{aligned} \quad (16)$$

as by assumption $\pi(\boldsymbol{\lambda}) \propto h(\boldsymbol{\lambda})$. \square

A.4 Proof of Proposition 3.3

By definition, the marginal density $\pi(\boldsymbol{\theta}) =$

$$\begin{aligned} \pi_u(\boldsymbol{\theta}) \int \frac{\pi(\lambda)}{h(\lambda)} \prod_{i=1}^p \mathbb{I}(d(\theta_i) > \lambda) d\lambda &= \pi_u(\boldsymbol{\theta}) \int \mathbb{I}(\lambda < d_{min}) \frac{\pi(\lambda)}{h(\lambda)} d\lambda = \\ & \pi_u(\boldsymbol{\theta}) P_\lambda(d_{min}) \int \frac{1}{h(\lambda)} \pi(\lambda \mid \lambda < d_{min}) d\lambda, \end{aligned} \quad (17)$$

where $P_\lambda(d_{min}) = P(\lambda < d_{min})$ is the cdf of λ evaluated at d_{min} . As $d_{min} \rightarrow 0$ we have that $\pi(\lambda \mid \lambda < d_{min})$ converges to a point mass at zero and hence the integral in the right hand side of (17) converges to $1/h(0) = 1$. To finish the proof of statement 1 notice that $P_\lambda(d_{min}) \propto d_{min} \pi(\lambda^*)$ by the Mean Value Theorem, as long as $P_\lambda(\cdot)$ is differentially and continuous around 0^+ , *i.e.* λ is a continuous random variable.

To prove statement 2, notice that $P_\lambda(d_{min}) \rightarrow 1$ as $d_{min} \rightarrow \infty$ and that the integral in the right hand side of (17) is $m(d_{min}) = E(1/h(\lambda) \mid \lambda < d_{min})$, which is increasing with d_{min} as $h(\lambda)$ is monotone decreasing in λ . Hence, $\pi(\boldsymbol{\theta}) \propto \pi_u(\boldsymbol{\theta}) m(d_{min})$ where $m(d_{min})$ increases as $d_{min} \rightarrow \infty$, *i.e.* $\pi(\boldsymbol{\theta})$ has tails at least as thick as $\pi_u(\boldsymbol{\theta})$. Furthermore, if $\int \frac{\pi(\lambda)}{h(\lambda)} < \infty$ the Monotone Converge Theorem applies and $m(d_{min})$ converges to a finite constant, *i.e.* $\pi(\boldsymbol{\theta}) \propto \pi_u(\boldsymbol{\theta})$. \square

A.5 Proof of Corollary 3.3

Analogously to the proof of Proposition 3.3, the marginal density can be written as $\pi(\boldsymbol{\theta}) = \pi_u(\boldsymbol{\theta}) \prod_{i=1}^p P_{\lambda_i}(d_i(\theta_i)) \times$

$$\int \dots \int \frac{1}{h(\boldsymbol{\lambda})} \pi(\boldsymbol{\lambda} \mid \lambda_1 < d_1(\theta_1), \dots, \lambda_p < d_p(\theta_p)) d\lambda_1 \dots d\lambda_p, \\ \pi_u(\boldsymbol{\theta}) E \left(h(\boldsymbol{\lambda})^{-1} \mid \forall \lambda_i < d_i(\theta_i) \right) \prod_{i=1}^p P_{\lambda_i}(d_i(\theta_i)), \quad (18)$$

where $h(\boldsymbol{\lambda})$ is a multivariate survival function and decreases as $d_i(\theta_i) \rightarrow 0$. Hence $E(h(\boldsymbol{\lambda})^{-1} \mid \forall \lambda_i < d_i(\theta_i))$ decreases as $d_i(\theta_i) \rightarrow 0$. To find the limit as $d_i(\theta_i) \rightarrow 0$ we note that the integral is bounded by the finite integral obtained plugging $d_i(\theta_i) = 1$ into the integrand. Hence, the Dominated Convergence Theorem applies, the integral converges to $E(h(\mathbf{0})) = 1$ and (18) to $\pi_u(\boldsymbol{\theta}) \prod_{i=1}^p P_{\lambda_i}(d_i(\theta_i))$. Since $\lambda_1, \dots, \lambda_p$ are continuous the Mean Value Theorem applies, so that $P_{\lambda_i}(d_i(\theta_i)) = d_i(\theta_i) \pi(\lambda_i^*)$ for some $\lambda_i^* \in (0, d_i(\theta_i))$. To prove statement 2, notice that $P_{\lambda_i}(d_i(\theta_i)) \rightarrow 1$ as $d_i(\theta_i) \rightarrow \infty$ and that $E(h(\boldsymbol{\lambda})^{-1} \mid \forall \lambda_i < d_i(\theta_i))$ increases as $d_i(\theta_i) \rightarrow \infty$. Hence, $\pi(\boldsymbol{\theta}) \propto \pi_u(\boldsymbol{\theta}) m(\boldsymbol{\theta})$ where $m(\boldsymbol{\theta})$ is increasing as $d_i(\theta_i) \rightarrow \infty$, which proves statement 2. Further, if $E(h(\boldsymbol{\lambda})^{-1}) < \infty$ the Monotone Convergence Theorem applies and $\pi(\boldsymbol{\theta}) \propto \pi_u(\boldsymbol{\theta})$. \square

A.6 Multivariate Normal sampling under outer rectangular truncation

The goal is to sample $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \Sigma) \mathbf{I}(\boldsymbol{\theta} \in T)$ with truncation region $T = \{\boldsymbol{\theta} : \theta_i < l_i \text{ or } \theta_i > u_i, i = 1, \dots, p\}$. We generalize the Gibbs sampling of Rodriguez-Yam et al. [2004] and importance sampling of Hajivassiliou [1993] and Keane [1993] to the non-convex region T .

Let $D = \text{chol}(\Sigma)$ be the Cholesky decomposition of Σ and $K = D^{-1}$ its inverse, so that $K\Sigma K' = KDD'K' = I$ is the identity matrix, and define $\boldsymbol{\alpha} = K\boldsymbol{\mu}$. The random variable $\mathbf{Z} = K\boldsymbol{\theta}$ follows a $N(\boldsymbol{\alpha}, I) \mathbf{I}(\mathbf{Z} \in S)$ distribution with truncation region S . Since $\boldsymbol{\theta} = K^{-1}\mathbf{Z} = D\mathbf{Z}$, denoting \mathbf{d}_i as the i^{th} row in D we obtain the truncation region $S = \{\mathbf{Z} : \mathbf{d}_i \cdot \mathbf{Z} \leq l_i \text{ or } \mathbf{d}_i \cdot \mathbf{Z} \geq u_i, i = 1, \dots, p\}$.

The full conditionals for Z_i given $Z_{(-i)} = (Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_p)$ needed for Gibbs sampling follow from straightforward algebra. Denote by d_{jk} the (j, k) element in D , then $Z_i \mid Z_{(-i)} \sim N(\alpha_i, 1)$ truncated so that either $d_{ji}Z_i \leq l_j - \sum_{k \neq i} d_{jk}Z_k$ or $d_{ji}Z_i \geq u_j - \sum_{k \neq i} d_{jk}Z_k$ hold simultaneously for $j = 1, \dots, p$. We now adapt the algorithm to address the fact that this truncation region is non-convex.

The region excluded from sampling can be written as $S_i^c = \bigcup_{j=1}^p (a_j, b_j)$, $a_j = (l_j - \sum_{k \neq i} d_{jk} Z_k) / d_{ji}$ when $d_{ji} > 0$ and $a_j = (u_j - \sum_{k \neq i} d_{jk} Z_k) / d_{ji}$ when $d_{ji} < 0$ (analogously for b_j). S_i^c as given is the union of possibly non-disjoint intervals, which complicates sampling. Fortunately, it can be expressed as a union of disjoint intervals $S_i = \bigcup_{j=1}^K (\tilde{a}_j, \tilde{b}_j)$ with the following algorithm. Suppose that l_i are sorted increasingly, set $\tilde{l}_1 = l_1$, $\tilde{u}_1 = u_1$ and $K = 1$. For $j = 2, \dots, p$ repeat the following two steps.

1. If $l_j > \tilde{u}_K$ set $K = K + 1$, $\tilde{l}_K = l_j$ and $\tilde{u}_K = u_j$, else if $l_j \leq \tilde{u}_K$ and $u_j \geq \tilde{u}_K$ set $\tilde{u}_K = u_j$.
2. Set $j = j + 1$.

Finally, because $(\tilde{l}_1, \tilde{u}_1), \dots, (\tilde{l}_K, \tilde{u}_K)$ are disjoint and increasing, we may draw a uniform number u in $(0, 1)$ excluding intervals $(\Phi(\tilde{l}_j), \Phi(\tilde{u}_j))$ and set $Z_i = \Phi^{-1}(u)$, where $\Phi(\cdot)$ is the inverse $\text{Normal}(\alpha_i, 1)$ cdf.

A.7 Monotonicity and inverse of iMOM prior penalty

Consider the product iMOM prior as given in (8). We first study the monotonicity of the penalty $d(\theta_i, \lambda)$, which for simplicity here we denote as $d(\theta)$, and then provide an algorithm to evaluate its inverse function. Equivalently, it is convenient to consider the log-penalty $\log(d(\theta)) =$

$$\frac{1}{2} (\log(\tau\tau_N) + 2\log(\phi) + \log(2)) - \log((\theta - \theta_0)^2) - \frac{\tau\phi}{(\theta - \theta_0)^2} + \frac{1}{2\tau_N\phi}(\theta - \theta_0)^2, \quad (19)$$

as its inverse uniquely determines the inverse of $d(\theta)$. Denoting $z = (\theta - \theta_0)^2$, (19) can be written as

$$g(z) = \frac{1}{2} (\log(\tau\tau_N) + 2\log(\phi) + \log(2)) - \log(z) - \frac{\tau\phi}{z} + \frac{1}{2\tau_N\phi}z. \quad (20)$$

To show the monotonicity of (20) we compute its derivative $g'(z) = -\frac{1}{z} + \frac{\tau\phi}{z^2} + \frac{1}{2\tau_N\phi}$ and show that it is positive for all z . Clearly, both when $z \rightarrow 0$ and $z \rightarrow \infty$ we have positive $g'(z)$. Hence we just need to see that there is some τ_N for which all roots of $g'(z)$ are imaginary, so that $g'(z) > 0$ for all z . Simple algebra shows that the roots of $g'(z)$ are $z = \tau_N\phi \pm \tau_N\phi\sqrt{1 - \frac{2\tau}{\tau_N}}$, so that for $\tau_N \leq 2\tau$ there are no real roots. Hence, for $\tau_N \leq 2\tau$ $g(z)$ is monotone increasing.

We now provide an algorithm to evaluate the inverse. That is, given a threshold t we seek z_0 such that $g(z_0) = t$. Our strategy is to obtain an initial guess from an approximation to $g(z)$ and then use continuity and monotonicity to bound the desired z_0 and conduct a linear interpolation based search. Inspecting the expression for $g(z)$ in (20) we see that the term $\log(z)$ is dominated by $\tau\phi/z$ when z approaches 0 and by $\frac{z}{2\tau_N\phi}$ when z is large. Hence, we approximate $g(z)$ by dropping the $\log(z)$ term, obtaining

$$g(z) \approx \frac{1}{2}(\log(\tau\tau_N) + 2\log(\phi) + \log(2)) - \frac{\tau\phi}{z} + \frac{1}{2\tau_N\phi}z. \quad (21)$$

Setting (21) equal to t and solving for z gives $z_0 = \tau_N\phi \left(-b + \sqrt{b^2 - 2\frac{\tau}{\tau_N}} \right)$ as an initial guess, where $b = \log(\tau\tau_N) + 2\log(\phi) + \log(2) - t$.

If $g(z_0) < t$ we set a lower bound $z_l = z_0$ and an upper bound z_u obtained by increasing z_0 by a factor of 2 until $g(z_0) > t$. Similarly, if $g(z_0) > t$ we set the upper bound $z_u = z_0$ and find a lower bound by successively dividing z_0 by a factor of 0.5. Once (z_l, z_u) are determined, we use a linear interpolation to update z_0 , evaluate $g(z_0)$ and update either z_l or z_u . The process continues until $|g(z_0) - t|$ is below some tolerance (we used 10^{-5}). In our experience the initial guess is often quite good and the algorithm converges in very few iterations.

Acknowledgements

This research was partially funded by the NIH grant R01 CA158113-01.

References

- M.J. Bayarri and G. Garcia-Donato. Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, 94:135–152, 2007.
- J.O. Berger and R.L Pericchi. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91:109–122, 1996.
- J.O. Berger and R.L. Pericchi. *Objective Bayesian methods for model selection: introduction and comparison*, volume 38, pages 135–207. Institute of Mathematical Statistics lecture notes - Monograph series, 2001.
- J.M. Bernardo. Integrated objective Bayesian estimation and hypothesis testing. In J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid,

- D. Heckerman, A.F.M Smith, and M. West, editors, *Bayesian Statistics 9 - Proceedings of the ninth Valencia international meeting*, pages 1–68. Oxford University Press, 2010.
- G. Consonni and L. La Rocca. On moment priors for Bayesian model choice with applications to directed acyclic graphs. In J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M Smith, and M. West, editors, *Bayesian Statistics 9 - Proceedings of the ninth Valencia international meeting*, pages 119–144. Oxford University Press, 2010.
- J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96: 1348–1360, 2001.
- J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–140, 2010.
- C. Fernández, E. Ley, and M.F.J. Steel. Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100:381–427, 2001.
- E. George, F. Liang, and X. Xu. From minimax shrinkage estimation to minimax shrinkage prediction. *Statistical Science*, 27:82–94, 2012.
- V.A. Hajivassiliou. Simulating normal rectangle probabilities and their derivatives: the effects of vectorization. *Econometrics*, 11:519–543, 1993.
- H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, England, third edition, 1961.
- V.E. Johnson and D. Rossell. Prior densities for default Bayesian hypothesis tests. *Journal of the Royal Statistical Society B*, 72:143–170, 2010.
- V.E. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 24(498):649–660, 2012.
- R.E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90:928–934, 1995.
- M.P. Keane. Simulation estimation for panel data models with limited dependent variables. *Econometrics*, 11:545–571, 1993.

- Irene Klugkist and Herbert Hoijtink. The bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, 51(12):6367 – 6379, 2007. ISSN 0167-9473.
- J.H. Kotecha and P.M. Djuric. Gibbs sampling approach for generation of truncated multivariate gaussian random variables. In *Proceedings, 1999 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1757–1760. IEEE Computer Society, 1999.
- F. Liang, R. Paulo, G. Molina, M.A. Clyde, and J.O. Berger. Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–423, 2008.
- E. Moreno, F. Bertolino, and W. Racugno. An intrinsic limiting procedure for model selection and hypotheses testing. *Journal of the American Statistical Association*, 93:1451–1460, 1998.
- A. O’Hagan. Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B*, 57:99–118, 1995.
- J.M. Pérez and J.O. Berger. Expected posterior prior distributions for model selection. *Biometrika*, 89:491–512, 2002.
- A. Raftery, J. Hoeting, C. Volinsky, I. Painter, and K.Y. Yeung. *BMA: Bayesian Model Averaging*, 2013. URL <http://CRAN.R-project.org/package=BMA>. R package version 3.16.2.3.
- G. Rodriguez-Yam, R.A. Davis, and L.L. Scharf. *Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression*. PhD thesis, Department of Statistics, Colorado State University, 2004.
- D. Rossell, D. Telesca, and V.E. Johnson. High-dimensional Bayesian classifiers using non-local priors. In *Proceedings of the Italian Classification and Data Analysis Society*. Springer (in press), 2012.
- J. Rousseau. Approximating interval hypothesis: p-values and Bayes factors. In J.M. Bernardo, M.J. Bayarri, J.O. Berger, and A.P. Dawid, editors, *Bayesian Statistics 8*, pages 417–452. Oxford University Press, 2010.
- M.D. Springer and W.E. Thompson. The distribution of products of beta, gamma and gaussian random variables. *SIAM Journal of Applied Mathematics*, 18(4):721–737, 1970.

- C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206, 1956.
- R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, B*, 58:267–288, 1996.
- I. Verdinelli and L. Wasserman. Bayes factors, nuisance parameters and imprecise tests. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 5*, pages 765–771. Oxford University Press, 1996.
- S. Wilhelm and B.G. Manjunath. tmvtnorm: a package for the truncated multivariate normal distribution. *The R Journal*, 2:25–29, 2010.
- A. Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques: essays in honor of Bruno de Finetti*, Amsterdam; New York, 1986. North-Holland/Elsevier.
- A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. In J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, editors, *Bayesian statistics: Proceedings of the first international meeting held in Valencia (Spain)*, volume 1. Valencia: University Press, 1980.
- A. Zellner and A. Siow. *Basic issues in econometrics*. University of Chicago Press, Chicago, 1984.